Análisis dataframes profiling

Contenido

Introducción

En este documento se hará un análisis al profiling de cada dataframe obtenido por la API TvMaze, donde se extrajeron los episodios emitidos del 1 al 31 de diciembre, para ello se uso graficas de correlación, estadística descriptiva, histogramas y graficas de valores nulos.

Los dataframes a analizar son:

- **Show_episodes :** Este dataframe contiene información correspondiente a los episodios de los shows registrados en la plataforma TVMaze, donde se encuentran columnas como por ejemplo: id, url, name, season etc
- **Shows:** Contiene información relacionada a los shows registrados en la plataforma TVMaze y tiene las siguientes columnas: show_id, show_url, show_name, show_genres etc
- **Web_channels:** Dataframe que dispone de la información del canal web donde el show es emitido, las columnas que contiene el dataframe son: webchannel_id, webchannel_name, webchannel_country_name etc
- **Networks:** el último dataframe a analizar es Networks el cuál muestra la cadena de televisión por donde se transmite el show, tiene información como: network_id, network_name, network_country_name

Análisis de dataframes y variables

Show_episodes:

Overview:

Overview | Alerts 37 | Reproduction

Dataset statistics

Number of variables	19
Number of observations	3082
Missing cells	15369
Missing cells (%)	26.2%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	457.6 KiB
Average record size in memory	152.0 B

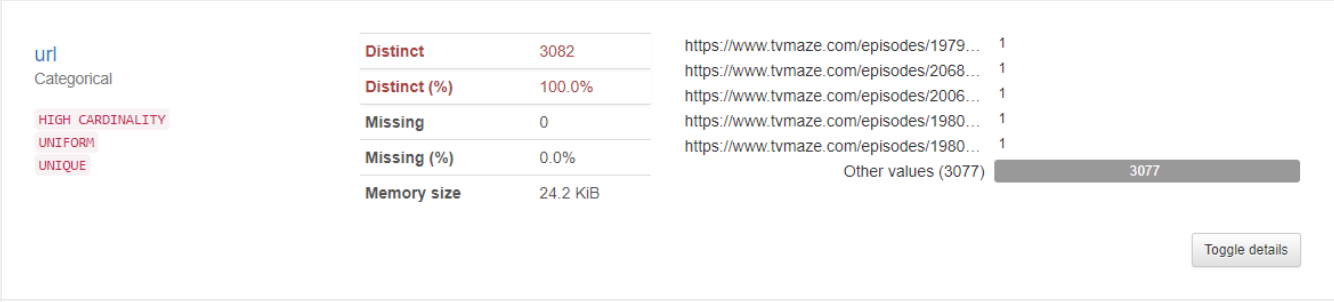
Variable types

Categorical	12
Numeric	6
Unsupported	1

el primer vistazo que tenemos de las variables es que tenemos 19 variables, es decir columnas de las cuales 12 son categóricas, 6 numéricas y 1 es indeterminada. Tenemos 3082 observaciones, sin embargo 15369 son celdas nulas y además no tenemos filas duplicadas:

Análisis variables:

Analizar las 19 variables puede saturar el propósito de este documento que es analizar cada dataframe, es por ello que revisaremos las variables más importantes:



La variable **url** (categórica) debido a que contiene el enlace a cada episodio todos sus registros son únicos

type

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	24.2 KiB

Toggle details

La variables **type** contiene 3 variables distintas y ninguna es nula (regular, significant_special,insignificant_special)

season

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	27	**Minimum**	1
Distinct (%)	0.9%	**Maximum**	2020
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	149.1966256	**Memory size**	24.2 KiB

Toggle details

Value	Count	Frequency (%)
1	1891	61.4%
2	354	11.5%
2020	224	7.3%
3	171	5.5%
4	156	5.1%
5	54	1.8%
7	41	1.3%
6	35	1.1%
18	23	0.7%
9	23	0.7%
Other values (17)	110	3.6%

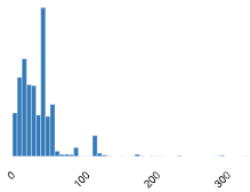
Variable **season** tiene 27 valores distintos, sin embargo se puede ver que las temporadas normalmente son números y no fechas, es por ello que la temporada 2020 hace que sea un valor diferente al consecutivo normal de la variable

runtime

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
MISSING

Distinct	112	Minimum	1
Distinct (%)	3.9%	Maximum	335
Missing	228	Zeros	0
Missing (%)	7.4%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	37.54344779	Memory size	24.2 KiB



Toggle details

Quantile statistics

Minimum	1
5-th percentile	7
Q1	19
median	31
Q3	45
95-th percentile	90
Maximum	335
Range	334
Interquartile range (IQR)	26

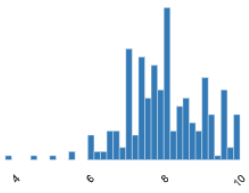
Variable **runtime** es la duración de cada episodio, podemos ver que la media es una duración de 37 minutos, sin embargo vemos que la distribución de los datos esta sesgado a la derecha , es por ello que los valores mas grandes “arrastra” o hace que aumente el valor de la media. Revisando los quartiles de las distribución podemos ver que el valor central es 31 minutos.

rating.average

Real number ($\mathbb{R}_{\geq 0}$)

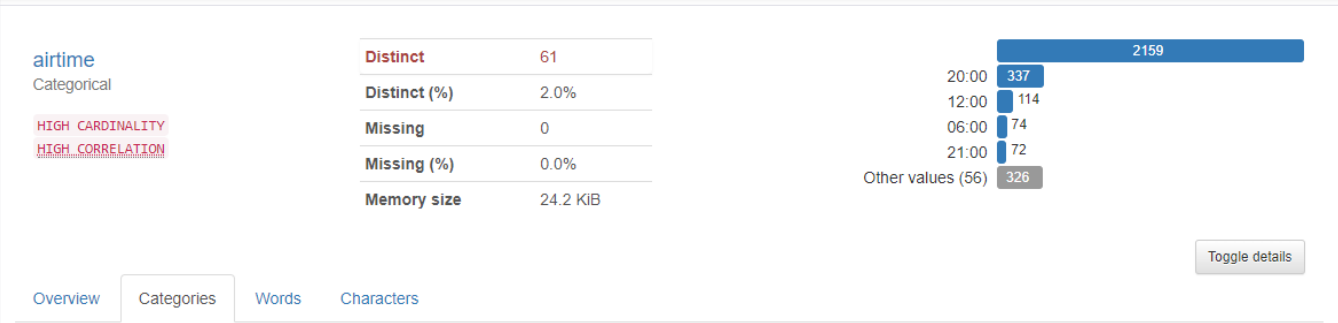
HIGH CORRELATION
MISSING

Distinct	37	Minimum	3.7
Distinct (%)	12.5%	Maximum	10
Missing	2786	Zeros	0
Missing (%)	90.4%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	7.97027027	Memory size	24.2 KiB



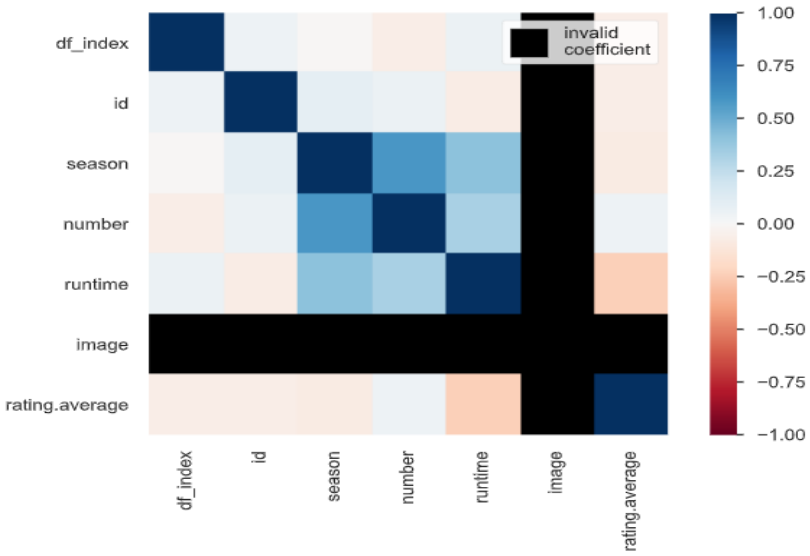
Toggle details

La variable **average_rating** nos muestra la calificación promedio en la página de los episodios emitidos en diciembre de 2020, sin embargo, vemos que el 90.4% de los datos son nulos, es decir no tienen calificación. Se obtiene una media de calificación de 7.9 sobre 10



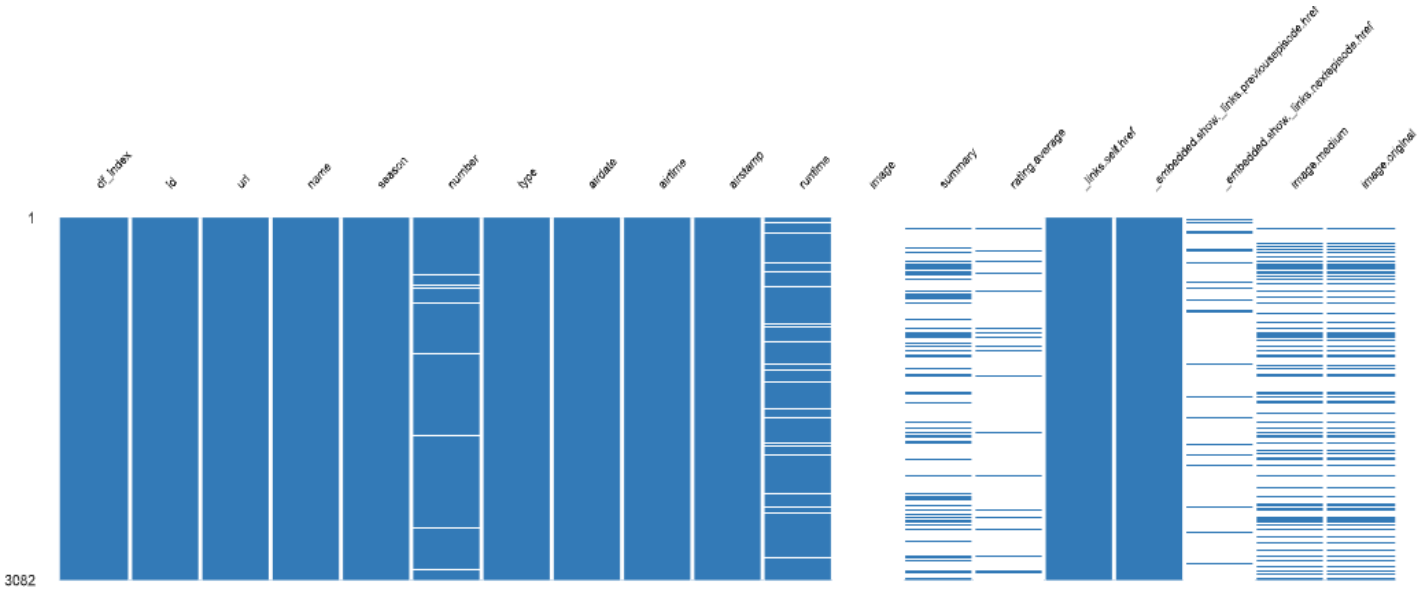
La última variable para analizar es **airtime** que es la hora en que el episodió se emitió. Vemos que no hay mucha información ya que 2159 episodios tienen valores en blanco, pero la mayoría de los episodios se emitió a las 20:00 siendo horario prime en la televisión.

Correlación:



En cuanto a la correlación entre las variables numéricas se puede observar que hay una baja correlación entre ellas, sin embargo la variable **season** tiene una correlación relativamente alta con las variables **runtime** y **number**, es decir entre mayor sea el número de la temporada así mismo aumenta el numero del episodio y la duración del episodio, sin embargo esto no explica causalidad

Valores nulos:



Se puede ver que las variables **image, summary, rating_average, next_episode, image_medium, image_original** tienen gran cantidad de valores nulos y pueden ser candidatos a ser eliminados o transformados en la limpieza del dataframe.

Shows:

Overview:

Overview | Alerts 55 | Reproduction

Dataset statistics

Number of variables	33
Number of observations	3082
Missing cells	32212
Missing cells (%)	31.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	794.7 KiB
Average record size in memory	264.0 B

Variable types

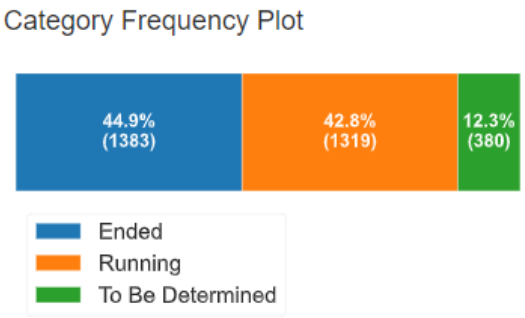
Categorical	17
Numeric	12
Unsupported	4

En cuanto al dataframe de shows tenemos 33 columnas de las cuales 17 son categóricas, 12 son números y 4 son indeterminadas. Como en el dataframe anterior hay 3082 observaciones y 32212 celdas nulas lo cual representa un 31.7% de celdas nulas con respecto al total de celdas, por otro lado no tenemos celdas duplicadas.

Análisis variables:

Value	Count	Frequency (%)
Scripted	1593	51.7%
Animation	385	12.5%
Documentary	328	10.6%
Reality	264	8.6%
Talk Show	242	7.9%
Variety	79	2.6%
Sports	79	2.6%
News	56	1.8%
Game Show	48	1.6%
Award Show	6	0.2%

La variable **show_type** posee 11 diferentes valores de los cuales ninguno es nulo, podemos ver que la mayoría de shows que emitieron capítulos en diciembre son “con guión”, 1593 shows representando un 51.7% sobre el total de shows.



_embedded.show.status

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory size	24.2 KiB

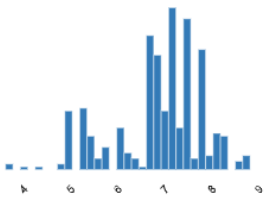
La variable **show_status** nos muestra tres valores distintos: **ended**, running y **to be determined**, las proporciones de los valores **ended**, running representan el 87.7% sobre el total, lo cual demuestra que la mayoría de shows están finalizadas o están en curso. Así mismo la variable no presente valores nulos.

_embedded.show.rating.a…

Real number (ℝ≥0)

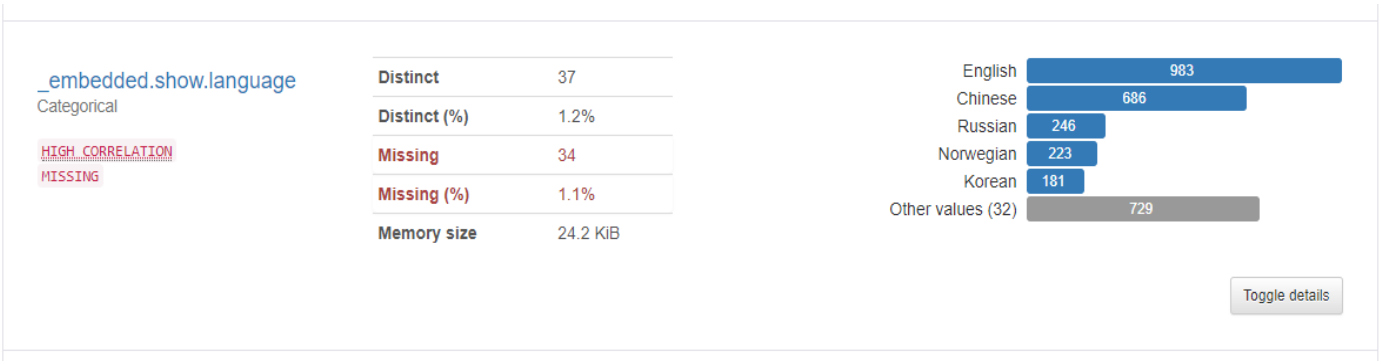
HIGH CORRELATION
MISSING

Distinct	33	Minimum	3.6
Distinct (%)	7.8%	Maximum	8.8
Missing	2661	Zeros	0
Missing (%)	86.3%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	6.864608076	Memory size	24.2 KiB

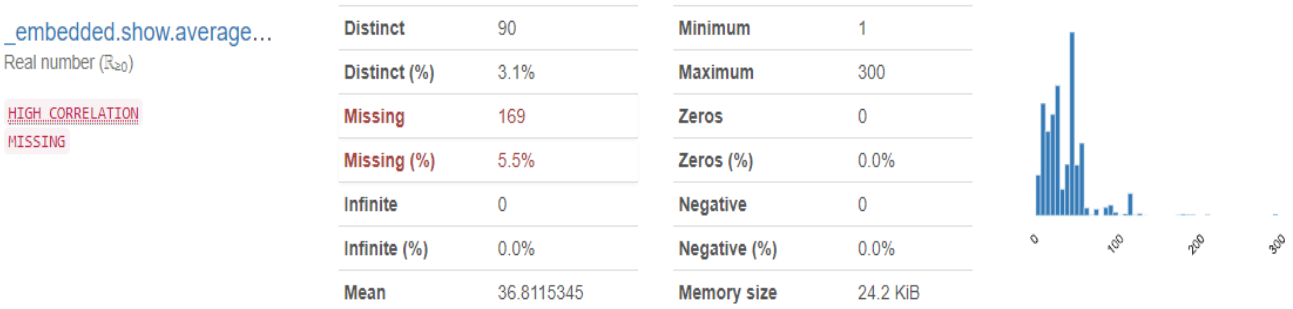


Toggle details

Hay un 86.3% de valores nulos para la variable **show_rating_average** representando un total de 2661 shows sin calificación en la plataforma, siendo este un número bastante grande, la distribución de los valores no tiene un sesgo marcado por lo que la media puede ser un buen indicador de medida de tendencia central con un 6.86 de calificación promedio sobre 10, siendo 8.8 la mas alta y 3.6 la más baja.

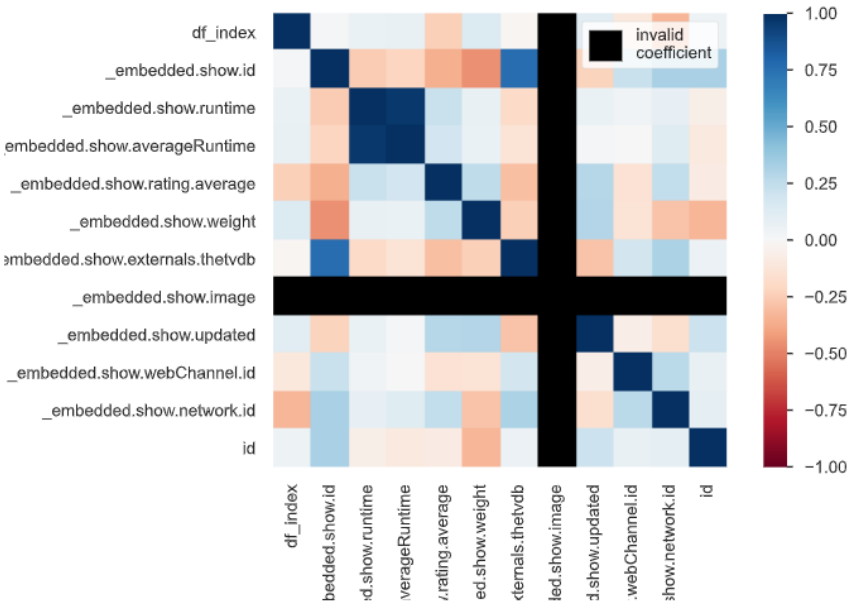


Se muestran 37 diferentes lenguajes de los shows en análisis, pero hay 34 shows que no tienen un lenguaje definido en la plataforma. Así mismo se puede decir de la mayoría de shows son en inglés representando un 31.9% , seguido por shows en idioma chino con un 22.3%



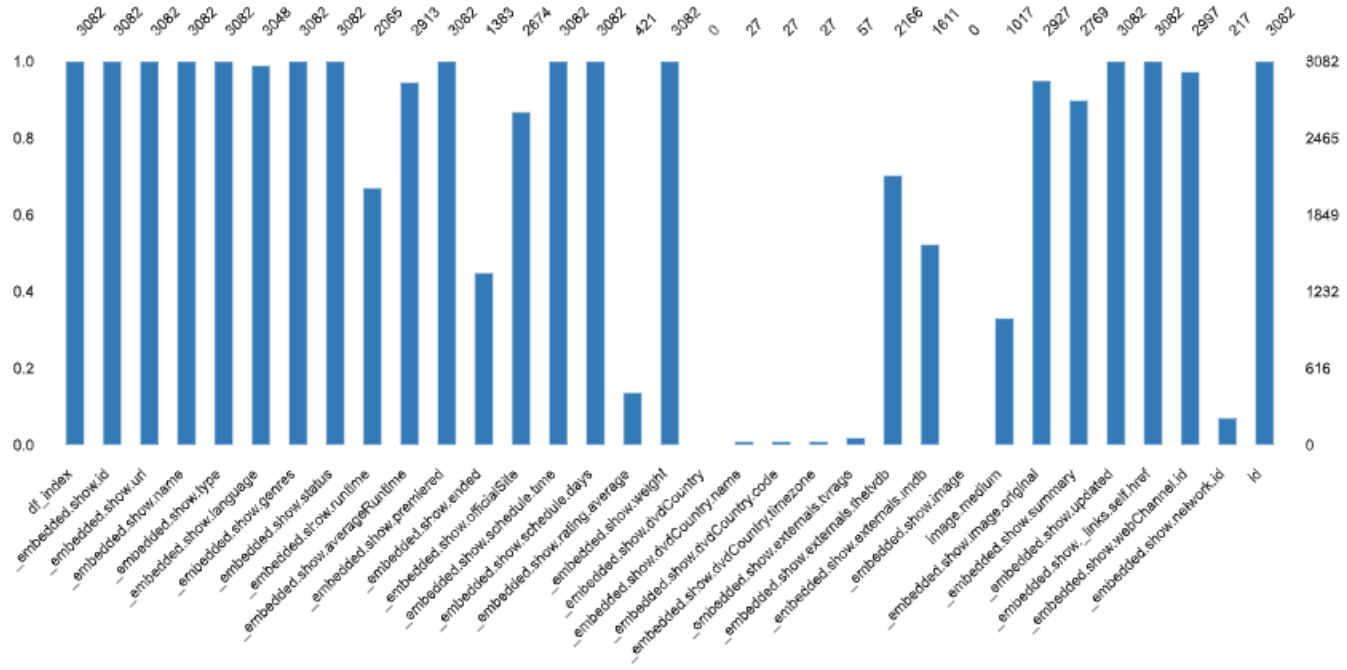
En cuanto a promedio de duración de los shows en promedio duran 36 minutos y 169 de ellos no tienen promedio de duración.

Correlación:



No hay correlaciones relevantes en las variables numéricas del dataframe shows, se muestra una correlación relativamente entre **thetvdb** y **show_id** sin embargo ambas variables son identificativos de los shows lo cual no tendría una relación entre ambas.

Valores nulos:



El grafico de valores nulos muestra que hay 11 variables con altos índices de valores nulos, siendo **show_dvdcountry, show_dvdcountry_name, show_dvdcountry_code, show_dvdcontru_timezone,**

show_Externals_tvrage, show_image y show_network_id las variables con los índices mas altos nulos, los cuales podrían modificarse o ser eliminados en la limpieza de los dataframe.

web_channels:

Overview:

Overview | Alerts 19 | Reproduction

Dataset statistics

Number of variables	9
Number of observations	3082
Missing cells	12325
Missing cells (%)	44.4%
Duplicate rows	509
Duplicate rows (%)	16.5%
Total size in memory	216.8 KiB
Average record size in memory	72.0 B

Variable types

Numeric	2
Categorical	5
Unsupported	2

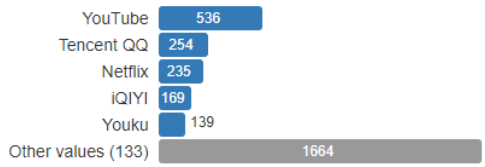
Grafico El dataframe web_channels tiene 9 diferentes variables, siendo 2 numéricas, 5 categóricas y 2 indeterminadas y además tiene 12325 celdas nulas representando un 44.4% sobre el total de celdas. Por otro lado, este dataframe a diferencia de los anteriores tiene 509 celdas duplicadas.

Análisis variables:

_embedded.show.webCh…
Categorical

HIGH CARDINALITY
MISSING

Distinct	138
Distinct (%)	4.6%
Missing	85
Missing (%)	2.8%
Memory size	24.2 KiB



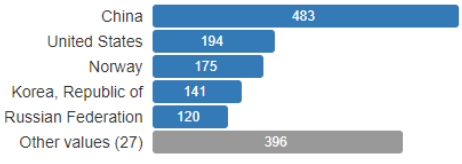
Toggle details

Los canales web de las distintas series son 138 y 85 shows no tienen web channel conocidos en la plataforma. Youtube es el canal para 536 shows (17.4%) seguido por Tencent QQ con una proporción de 8.2% y Netflix con 7.6%.

_embedded.show.webCh…
Categorical

HIGH CORRELATION
MISSING

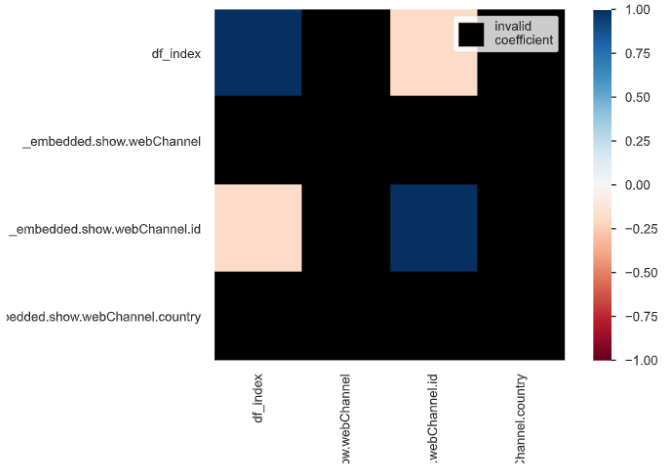
Distinct	32
Distinct (%)	2.1%
Missing	1573
Missing (%)	51.0%
Memory size	24.2 KiB



Toggle details

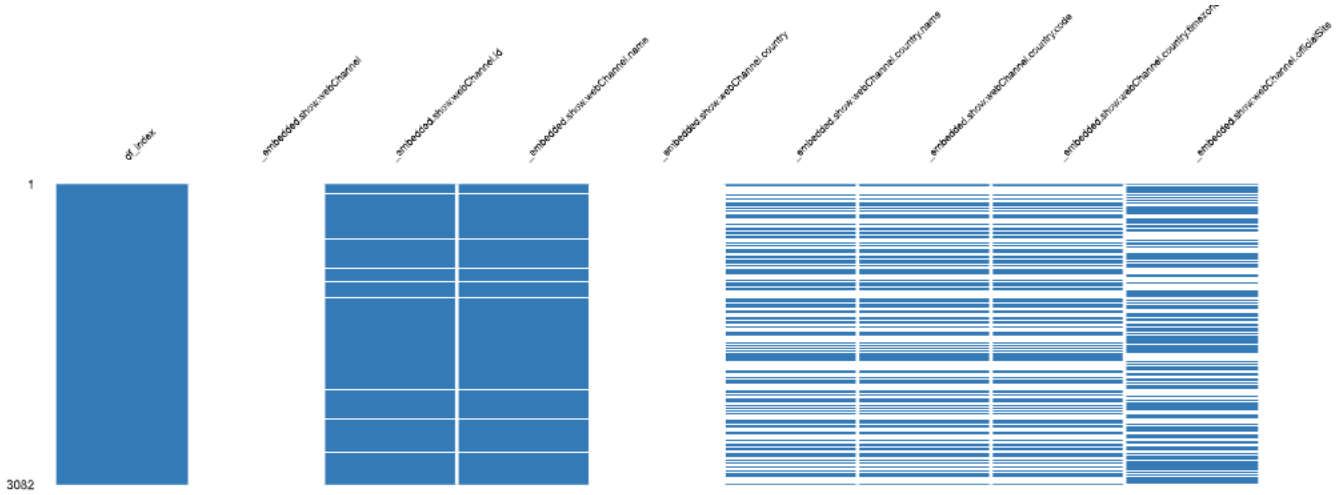
Los canales web de los shows de China representan un 15.7% sobre el total, seguido por Estados Unidos con un 6.3% y Noruega con 5.7%

Correlación:



No existe correlación entre las variables del dataframe web_channels

Valores nulos:

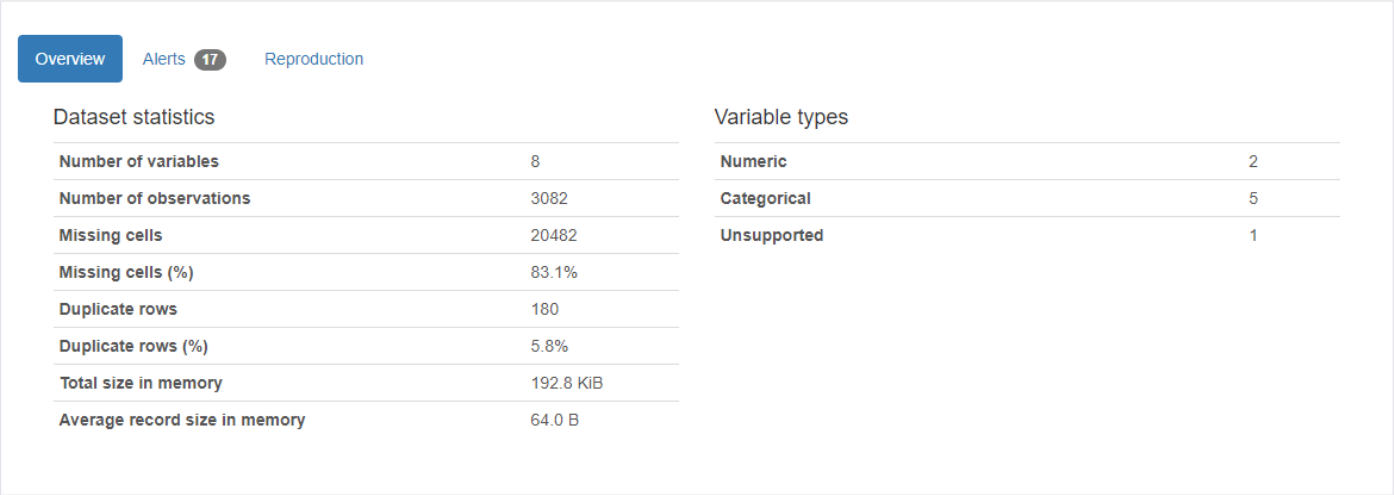


Las variables web channel y **channel_country** son totalmente nulas, además hay algunos ids y nombres de canales que son nulos, así mismo hay varios valores nulos **channel_country_name**, **channel_country_code y channel_country_timezone** y en menor medida **channel_officialsite**.

Este dataframe tiene muchos valores nulos lo cual puede representar un inconveniente para relacionar la tabla con los shows.

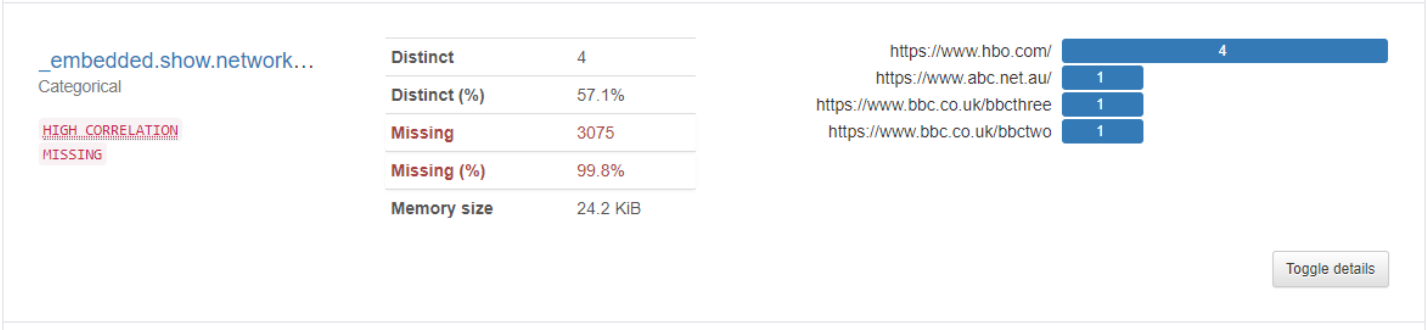
network_channels:

Overview:

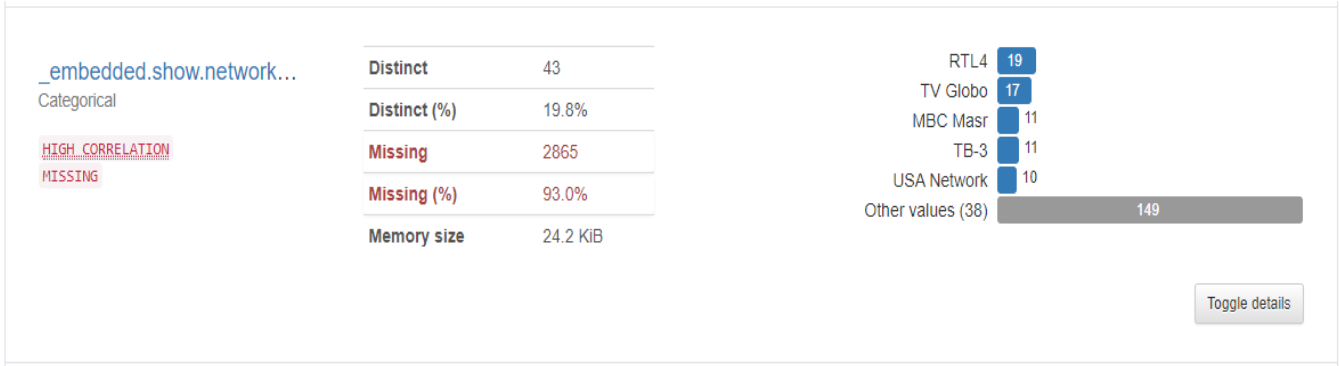


Tenemos 8 variables para el dataframe network_channels con un alto porcentaje de 83.1 de valores nulos con 180 filas duplicadas representando 5.8%. 2 variables son numéricas, 5 categóricas y 1 indeterminada.

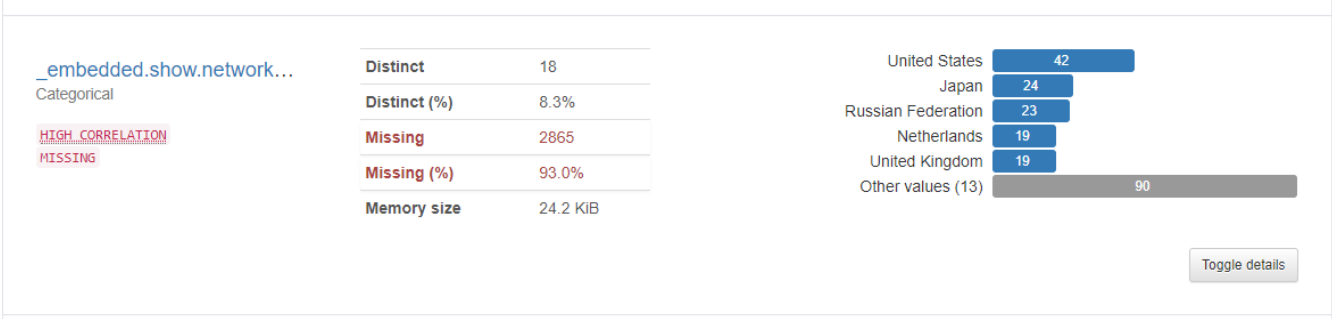
Análisis variables:



Casi el 100% de los valores de la variable **network_offical_site** son nulos lo cual hace que esta variable esté descartada

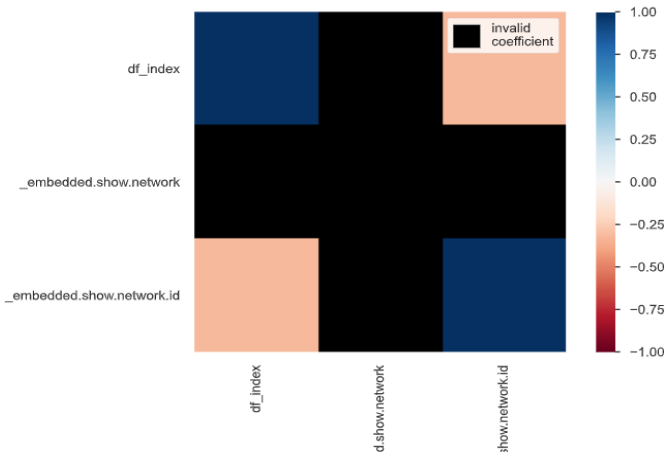


Hay 43 diferentes valores para la variable **network_name** lo cual muestra las diferentes cadenas de televisión de los shows emitidos. Sin embargo, el 93% de los valores son nulos lo cual demuestra que este dataframe no es significativo en cuanto a calidad de datos.



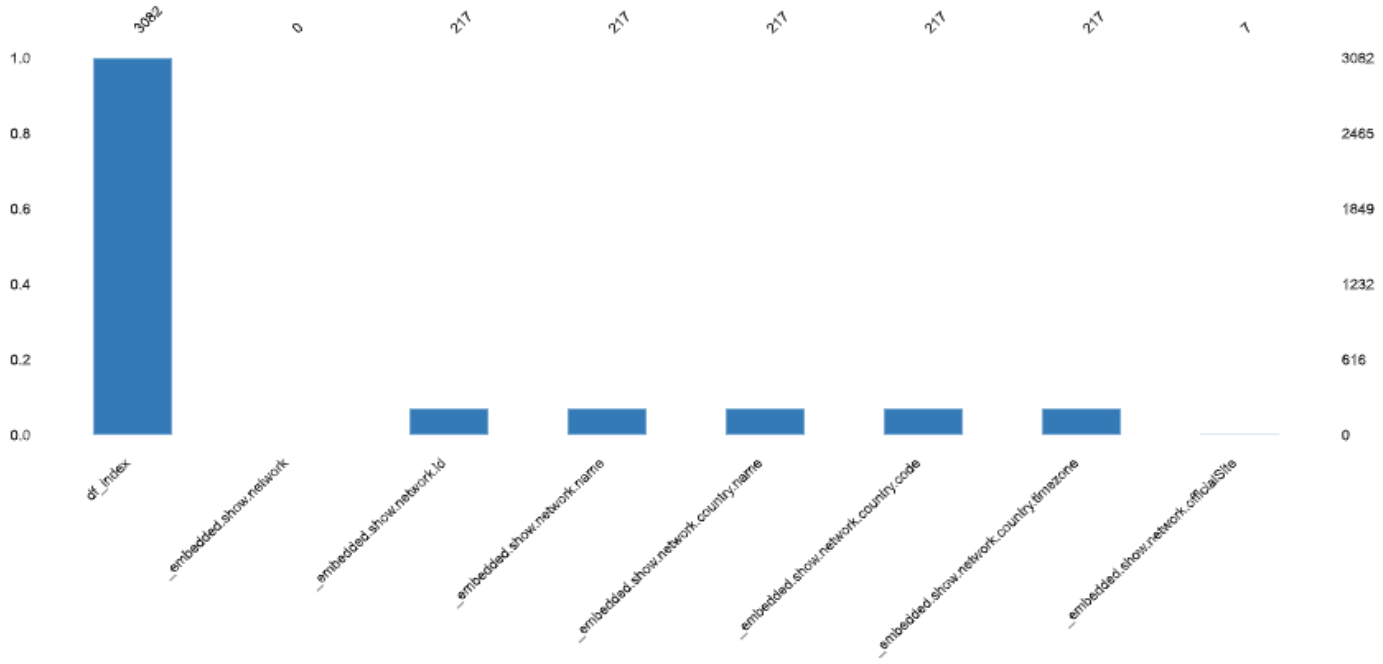
El 93% de los valores son nulos para la variable **network_country_name** lo que marca la tendencia de que las variables en tiene en su mayoría valores nulos

Correlación:



No existe correlación entre las variables del dataframe

Valores nulos:



Como veíamos en el análisis de las variables la mayoría de ellas tienen valores nulos indicando que el dataframe no es significativo para el análisis de los shows.

Conclusiones

Los cuatro dataframes tienen valores nulos en sus variables sin embargo el dataframe network_channels posee un aproximado de 83% de celdas nulas lo cual representa un gran porcentaje con respecto al total haciendo que este no sea significativo para el análisis de los shows, sin embargo se borraran solamente las columnas que no sean significativas para el análisis mas no se borraran los dataframe. Además, no existe una correlación notable entre todas las variables analizadas.